KUOPION YLIOPISTON JULKAISUJA H. INFORMAATIOTEKNOLOGIA JA KAUPPATIETEET 8
KUOPIO UNIVERSITY PUBLICATIONS H. BUSINESS AND INFORMATION TECHNOLOGY 8

NIINA PÄIVINEN

Scale-free Clustering

A Quest for the Hidden Knowledge

Doctoral dissertation

To be presented by permission of the Faculty of Business and Information Technology of the University of Kuopio for public examination in Auditorium Sky 1, Microtower building, University of Kuopio, on Friday 30th March 2007, at 12 noon

Department of Computer Science
University of Kuopio

KUOPION YLIOPISTO

KUOPIO 2007

Distributor:
Kuopio University Library
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 17 163 430
Fax +358 17 163 410
www.uku.fi/kirjasto/julkaisutoiminta/julkmyyn.html

Series Editors:
Professor Markku Nihtilä, D.Sc.
Department of Mathematics and Statistics

Assistant Professor Mika Pasanen, D.Sc.
Department of Business and Management

Author's address:
Department of Computer Science
University of Kuopio
P.O. Box. 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 17 162 172
Fax +358 17 162 595
E-mail: niina.paivinen@cs.uku.fi

Supervisors:
Docent Tapio Grönfors, Ph.D.
Department of Computer Science
University of Kuopio

Research director Seppo Lammi, Ph.D.
Department of Computer Science
University of Kuopio

Reviewers:
Professor Erkki Mäkinen, Ph.D.
Department of Computer Sciences
University of Tampere

Jarno M.A. Tanskanen, D.Sc.
Ragnar Granit Institute
Tampere University of Technology

Opponent:
Professor Pasi Fränti, Ph.D.
Department of Computer Science and Statistics
University of Joensuu

ISBN 978-951-781-987-9
ISBN 978-951-27-0106-3 (PDF)
ISSN 1459-7586

Kopijyvä
Kuopio 2007
Finland

Päivinen, Niina. Scale-free Clustering: A Quest for the Hidden Knowledge. Kuopio University Publications H. Business and Information Technology 8. 2007. 57 p.
ISBN 978-951-781-987-9
ISBN 978-951-27-0106-3 (PDF)
ISSN 1459-7586

ABSTRACT

Clustering is a data mining procedure in which information is extracted without supervision from a dataset. The goal of clustering is to partition the dataset into clusters in such a way that the objects in the same cluster are similar with each other, whereas objects in different clusters are different from each other.

The most well-known approaches to the clustering problem are presented along with the question of clustering validity. When comparing these methods with the design principles for clustering methods, it can be seen that all the principles are not always met. Each method has also its own drawbacks, and there is no method suitable for all situations.

Three complex network models, random, small-world and scale-free, are introduced, and the scale-free model is studied more closely. The possible uses of graph theory in clustering are presented, and the scale-free model is brought to clustering in the form of scale-free minimum spanning tree (SFMST). The structure is used as a way of obtaining clustering.

The new clustering method proposed in this thesis, the SFMST clustering, has been created with the design principles in mind: the method needs only one control parameter; it can find clusters with different shapes, sizes and densities; the optimal number of clusters can be detected automatically; the method is not too demanding computationally. The method has been tested and evaluated with different datasets, both real and artificial. The results have been found promising and the method can be said to be a potential clustering method.

Universal Decimal Classification: 001.82, 025.44/.47, 004.82, 004.421, 004.62, 519.17
Inspec Thesaurus: knowledge acquisition; data mining; pattern clustering; pattern classification; network analysis; graph theory

Acknowledgments

I want to thank the supervisors, professor Tapio Grönfors and research director Seppo Lammi, for their guidance. Special thanks go to Dr. Grönfors; this thesis would not exist without him. He gave me the idea of taking a closer look at scale-free networks, and that encouraged me to write an algorithm for constructing a minimum spanning tree with a scale-free structure. I also want to thank the reviewers, professor Erkki Mäkinen and doctor Jarno M. A. Tanskanen, for giving helpful comments from diverse areas of the manuscript.

The work behind this thesis has been carried out in Department of Computer Science at the University of Kuopio, where the author has been an assistant and an acting senior assistant. The work has been funded completely by University of Kuopio. I would like to thank the entire staff of Department of Computer Science for making the working environment pleasant.

Finally, thanks to my family and friends for encouragement and support, especially to Antti and Aada for their love and patience, and for taking me to walks.

List of original publications

This study is based on the following publications referred to in the text by their Roman numerals.

I Niina Päivinen and Tapio Grönfors. Minimum spanning tree clustering of EEG signals. In *Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG 2004)*, June 9–11, Espoo, Finland, pages 149–152, 2004. IEEE.

II Niina Päivinen. Clustering with a minimum spanning tree of scale-free-like structure. *Pattern Recognition Letters*, 26(7):921–930, 2005.

III Niina Päivinen and Tapio Grönfors. Modifying the scale-free clustering method. In *Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control & Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)*, 28–30 November 2005, Vienna, Austria, volume II, pages 477–483, Los Alamitos, California, 2006. IEEE.

IV Niina Päivinen and Tapio Grönfors. Finding the optimal number of clusters from artificial datasets. In *Proceedings of 2006 IEEE International Conference on Computational Cybernetics (ICCC 2006)*, 20–22 August 2006, Tallinn, Estonia, pages 155–160, 2006. IEEE.

V Niina Päivinen. A fast clustering method using a scale-free minimum spanning tree. Manuscript submitted to *Information Processing Letters*.

Contents

Chapter 1

Introduction

Clustering is a core task in data mining [HK01], which is becoming more and more important since the digital data storage capacity is growing faster than expected [FU02]. There are more and more data available, but the processing techniques have not developed correspondingly. Data tombs — data stores that are too big to handle and are in practice write-only — have emerged. Data tombs are often never accessed after they have been created, and the knowledge in them falls into oblivion. This kind of resource wasting is clearly not desired; the data tombs might contain crucial information hidden in the vast amounts of data. Thus, the data miner has ”a quest for the hidden knowledge” before him. His aim is to seek the information from the data tomb, and one way to attain the goal is the usage of clustering methods.

Although there are many clustering methods already available, it is not known in advance which method performs best for a given application [DHS01]. Clustering is an empirical research area in which there is still much to do, and the development of novel methods is justified.

The main principle of clustering is that the members of a cluster are similar to each other. One way of measuring this similarity among mathematicians is the well-known Erdös number. Paul Erdös was an award-winning, productive mathematician with the largest number of different co-authors in mathematics. A researcher's Erdös number measures the length of the chain of authors from Erdös to the researcher; the co-authors of Erdös (509 researchers) have Erdös number 1, the researchers being co-authors with the ones with Erdös number 1 get Erdös number 2, and so on. If the researchers collaborated with each other are connected with a line, the result is a network of people. There are cliques (groups of researchers for which there is a connecting edge for each pair of researchers) in this graph as well as researchers having strong influence to their

collaborators (highly connected researchers). The Erdös numbers illustrate the curiosity about how we do in the research world, and how closely connected we are, socially or in some other way. Illustrating social connections with networks can clearly reveal our status. [BM00, Gro05, New04]

To gain a new insight into the clustering problem, this study takes advantage of the network structure described above. The highly connected researchers can be interpreted as cluster centers, and the collaborators of the researcher obviously belong to the same cluster with him.

In 1999, Albert-László Barabási found with his colleagues that the architecture of world-wide Web was not as expected [BA99]. They assumed that the link distribution of Web sites would follow a Gaussian distribution, and the average number of links, or the ”scale” of the network, could be determined. In contrast, it was revealed that the Web is dominated by a few highly-connected sites, whereas the most sites have only a few links. There was no definite scale to be seen — thus the name ”scale-free” network arose. Since then, the same kind of network structure has been found in diverse fields such as economy, molecular biology, quantum physics and social networks, and this phenomenon is also becoming popular with the general public [Coh02].

1.1 The aims of the study

The goal of the research was to study the possible usage of new network models in clustering. During the study, a clustering method based on graph theory and scale-free structures was developed and evaluated. To the best of our knowledge, the usage of scale-free graphs in clustering is a novel idea. To achieve a functional clustering method, a few design principles must be considered. These include

- the method does not use a lot of numerical parameters whose values have to be provided by the user;
- the method is capable of finding clusters with different sizes, shapes and densities;
- the optimal number of clusters is detected automatically;
- the method should not be computationally demanding.

If existing clustering methods are viewed with these principles in mind, it can be noticed that the principles are not always met. It seems that usually one clustering method takes into account one principle but leaves the other ones

intact. This causes problems in selecting a method for a certain application: if the user is not familiar with different methods, he may base the selection on misleading facts. The clustering method developed during this study fulfills the principles and its drawbacks are known, and these facts together make it a potential method.

1.2 The organization of the thesis

The thesis is organized as follows. Chapter 2 discusses clustering in general. The basic principles of clustering are reviewed as well as the most well-known approaches to the problem. In addition, the question about clustering validity, including the quality of the clustering and the number of clusters, is discussed.

In Chapter 3, networks are introduced. Three models for complex networks, namely random, small-world and scale-free, are presented. The scale-free model and its most interesting properties are studied more closely.

Chapter 4 merges clustering with the networks, and deals with graph-theoretic clustering. Methods based on the minimum spanning tree along with other graph-theoretic clustering methods are reviewed, and applications are also presented.

Chapter 5 contains a summary of the publications **I**–**V**. The main results of each publication are presented.

Finally, in Chapter 6, conclusions, discussion and ideas for future work are presented.

Chapter 2

On clustering

Pattern recognition is a broad concept consisting of different methods for finding and classifying patterns from a dataset. A pattern is some entity of interest which can be represented, for example, as a feature vector containing feature values, or some measurable properties of the pattern. Examples of patterns might include DNA sequences, fingerprints, spoken words, or hand-written numerals. The approach of using feature vectors, namely the statistical approach to pattern recognition, is most used in practice. There are many different approaches in statistical pattern recognition depending on what is known about the classes of the dataset. If the conditional densities of the classes are not known, they can be learned from the dataset; if the learning process is done without supervision, the process is called clustering.

The goal of clustering process is to classify the data points into classes, clusters, in such a way that the elements in the same cluster are more similar to each other than with an element from a different cluster. The quality of the clustering can be measured with a classification error function. Applications of clustering include image segmentation (partitioning an input image into homogeneous regions), object and character recognition, information retrieval (automatic storage and retrieval of documents), data mining (extracting meaningful information from data stores), and investigating the functions of genes. [DHS01, JMF99, JDM00, KLV98, TK03, XWI05]

The first essential question arising when starting cluster analysis on a dataset is the selection of the features to represent the data points. Features can be divided into two types, quantitative and qualitative [JMF99]. Quantitative features include continuous, discrete and interval values, whereas nominal and ordinal values are qualitative. The feature extraction methods include, for example, principal component analysis, linear discriminant analysis, multidi-

mensional scaling, self-organizing map and other projection, decomposition or transform methods [JDM00, TK03]. A feature extraction method based on the concept of mutual information has also been proposed [FIP98]. The feature extraction problem has not been widely discussed in the literature, but it has been shown that it might be beneficial to use a combination of features based on different ideas in the same classification problem [PLP+05]. The dataset might already contain measurements suitable to be the features — if this is the case, normalization of the features may still be needed before the actual clustering in order to avoid the dominance of some features over others. In addition, if some features are strongly correlated with each other, the results can become skewed. Thus, not all the available features are necessarily needed. The feature selection problem is widely discussed in the literature, and there are many different methods to select the features [JDM00, TK03].

The second essential question deals with the similarity measure. Since the goal is to bundle together data points similar with each other, the first thing to do is to define this similarity. The most natural selection is a distance measure such as the Euclidean metric [XWI05]. However, if all the features are not continuous-valued, a metric may not be the best choice. For nominal features, different matching coefficients might be used [Fin05, JD88], and new distance functions being able to handle nominal features, continuous features or both have also been presented [WM97]. There have also been studies concerning the classification of time series data into stationary and non-stationary, and different metrics have been suggested to be used in this case [CCP06]. They might also be usable in other classification problems. In addition, a similarity measure based on mutual information to be used with hierarchical clustering [Koj04] as well as a metric based on Fisher information matrix [KSP01] have also been presented. The latter is a metric which is learned from the dataset, and it has been used to predict a bankruptcy from an enterprise's financial data. Wong et al. have also considered the possibility of learning the similarity measure from the dataset, and have formulated a new clustering algorithm based on the idea [WCS01].

If the chosen similarity measure is a metric, the results for search problem in metric spaces [CNBYM01] and for approximate searching [AMN+98] can be used in clustering problems. It has been claimed that in higher-dimensional spaces (with 10–15 dimensions) the concept of "nearest neighbor" is not any more reasonable [BGRS99]. This may have consequences in clustering procedures also.

2.1 On clustering methods

The selection of an appropriate clustering method for a given problem is an even more difficult task than selecting the similarity measure. There are lots of methods available, each with different characteristics. The clustering method can be either hard or fuzzy, depending on whether a data point is allowed to belong to more than one cluster, with a definite degree of membership, at the same time. In addition, a method can be called hierarchical or partitional; hierarchical methods produce a nested sequence of partitions whereas partitional methods produce only one partition. There are methods based on squared error (such as the k-means), probability density function estimation, graph theory, combinatorial search techniques, neural networks, and others. Different strategies have also been proposed for sequential or high-dimensional data and large datasets. [JMF99, XWI05]

A very popular clustering method, the k-means method, is also the best-known squared error-based clustering algorithm. The method begins by initializing k cluster centers, and then proceeds by assigning data points to the center nearest to them, re-calculates the cluster centers, and assigns the points again. The process ends when there is no change in the cluster centers. The method is simple and easy to understand, but it has its drawbacks. The initial cluster centers and the number of clusters have to be given to the algorithm. The method is iterative and there is no guarantee that it converges to a global optimum, and the method is sensitive to outliers and noise. [DHS01, XWI05] Furthermore, the k-means method can only detect hyperspherical clusters (if Euclidean distance is used) [JDM00]. There is still ongoing research aiming to improve the k-means method. For example, an efficient exact algorithm for finding optimal k centers has been given [AP02], as well as a way to estimate the number of clusters present in the dataset with statistical methods during the k-means clustering procedure [PM00].

Since clusters of different size, shape and density create problems in clustering, many solutions have been offered. An example of them is a method using the information about nearest neighbors of the data points in defining the similarity [ESK03]. The method uses core points to represent the clusters, and it has been shown to outperform k-means. It is also possible to use cluster skeletons instead of cluster centers [YCC00]. This approach can handle clusters with different shapes correctly.

Another solution is based on neural networks. For humans, visual pattern recognition is something we do easily every day, for computers the same problem is difficult. Thus, the human brain has been proposed as a model to be

used in pattern recognition [Fuk88]. An artificial neural network (ANN) is a structure modeling the functionality of human brain. ANNs have been used in clustering problems since they are naturally nonlinear, massively parallel distributed systems being able to learn and generalize [Hay94]. It has also been claimed that a pulse-coupled neural network can be used to model pattern recognition by the human brain [Hak05].

Estimation of probability density functions behind the data to be clustered is a way to overcome problems with overlapping, varyingly sized and shaped clusters. In model-based clustering [BR93a, BR93b] it is assumed that the dataset is generated by a mixture of probability distributions in which each component represents a different cluster. The actual clustering can be found by a combination of hierarchical clustering and the expectation-maximization (EM) algorithm [FR98]. The number of components in the model, i.e., the number of clusters, can be determined using the Bayesian information criterion or approximate Bayes factors [DR98]. The noise points can be represented by a spatial Poisson process. Actually, in the clustering methods based on the mean-square error minimization, such as the k-means method, one is fitting Gaussian mixtures to the data [BR93a].

The standard EM algorithm needs initialization to work properly in mixture fitting. To avoid the initialization, a method for learning mixture models without supervision has been proposed [FJ02a]. The probability density function can also be estimated by using Parzen windows [HBV96], or using Gaussian mixture models, with which it is possible to derive many clustering criteria related with the shapes and densities of the clusters [CG95].

There has been discussion in the literature about classifier combination for better recognition accuracy [KHDM98, JDM00]. The idea of evidence accumulation -based clustering, or combining the results of multiple clusterings using a split-and-merge approach, has also been presented [FJ02b].

2.2 On clustering validity

If a dataset is processed using a clustering method, the outcome is a clustering, meaning, a collection of subsets of the original dataset. It is entirely possible to cluster a dataset not containing any intrinsic clusters, and to get a result. How can the number of clusters in a dataset be found out, and how can the quality of the clustering result be measured? These questions are difficult, since the definition of a cluster is not always apparent. The criterion ”similar points belong to the same cluster” might seem exact, but the similarity can be defined

in many different justified ways. Clustering is ultimately beholder-dependent [EC02]. An example study of evaluating the performance of different classifiers using minimum classification error as the criterion has been presented by Sohn [Soh99].

It has been claimed that artificial datasets are essential when evaluating the performance of data mining procedures, since real datasets may contain structural regularities whose nature is not known [SW99]. In addition, the degree of difficulty of the dataset might be needed to be varied during the performance evaluation, for example, by adding error perturbation to the data [Mil80]. This is clearly not possible with real datasets. Procedures for generating artificial dataset have been presented in the literature [Mil85, SW99].

2.2.1 Number of clusters

When talking about cluster validity, the number of clusters is obviously an important issue. The problem has been studied widely, and many attempts to estimate the number of clusters from a given datasets have been made. These include the construction of certain indices, optimization of a criterion function, and other heuristic approaches. In addition, some clustering methods are able to adjust the number of clusters during the processing. Sometimes it might also be possible to project the dataset into two-dimensional space, and see the needed number of clusters from the visualization. [XWI05] An example study in which the number of clusters are determined graphically can be found in the application area of microarray gene expression data [Bic03]. There exists also the Visual Assessment of Tendency (VAT) algorithm for studying the clustering tendency of a dataset [HBH05].

The indices, or stopping rules, are usually based on cluster cohesion and isolation. These indices can be either method-dependent or independent. Method-independent indices have been compared with each other in the case of artificial datasets [MC85, DDW02]. An index based on the comparison of points inside a cluster with the points between the clusters has also been proposed [WDRP02].

Defining the number of clusters statistically includes methods such as the gap statistic [TWH01], a simulated annealing clustering based method [LF01], cluster isolation criterion [FL03], cluster stability [BBM06], and Rand's statistic [CDW06].

A collection of methods for estimating the number of clusters in a dataset, and a new prediction-based resampling method are compared with each other in the case of gene expression microarray data in an article by Dudoit and

Fridlyand [DF02]. Hardy presents methods based on hypervolume criterion along with a few statistical criteria, and studies their performance with artificial datasets [Har96].

Minimizing the regularized cost function has also been presented as a way to find the number of clusters [KP99], as well as a separation index aiming to measure the gaps between the clusters [QJ06].

If the problem is the definition of the number of clusters in mean square error sense, or the number of Gaussians in a finite Gaussian mixture by the EM algorithm, the answer can be given by the Bayesian Ying–Yang machine [Xu96, Xu97, GCL02].

Using influence zones has been proposed as a way to determine the number of clusters without constructing the clustering itself [HBV96, HBV01]. The method is suitable to one-, two- or three-dimensional data; higher-dimensional data has to have its dimensionality reduced before the method can be applied. In this approach, the probability density function of the data is estimated. The method is applied to the dataset with different values of a certain parameter, and the number of clusters can be selected according to the behavior of this parameter.

Genetic algorithm-based approaches to finding the number of clusters include, for example, an algorithm automatically evolving the number of clusters during the clustering process [BM02], a genetic algorithm maximizing an objective function based on the average silhouette width criterion to find the right number of clusters [HE03], and a genetic algorithm being able to handle non-spherical clusters and at the same time finding the proper number of clusters [TY00].

A fuzzy approach to the problem with the number of clusters has also been presented [FK96]. In this approach, cluster prototypes are being estimated from noisy dataset. The Cluster centers can also be found with fuzzy ants: the ants move the cluster centers in the feature space, and the centers are then evaluated and passed to fuzzy c-means algorithm [KH04].

A cluster analysis algorithm which does not produce the explicit clustering, but an ordering of the dataset, has been proposed by Ankerst et al. [ABKS99]. The method is claimed to reveal the intrinsic clustering structure of the dataset, including distribution and correlation of the data.

2.2.2 Validity

Cluster validation includes methods that evaluate the results of cluster analysis quantitatively. There are three different kinds of testing criteria: external,

internal, and relative. External criteria measure performance by comparing the clustering result with the information known a priori, internal criteria consider the quality of the partition in the viewpoint of the dissimilarity matrix, and relative criteria compare two clusterings with each other. Internal and external criteria are based on statistical testing, whereas relative criteria do not involve statistical tests and thus, they are computationally less demanding than internal and external criteria. [JD88, XWI05, HBV02a, HBV02b]

Maybe the simplest possible performance measures can be derived from the quantities of true positive, true negative, false positive, and false negative instances. These quantities can only be used in the case of binary classification problem. It is also possible to use different distance measures and derivatives of them, or correlation coefficients, calculated between vectors of the known classes and the ones given by a classifier. [BBC$^+$00]

Information theory provides possible measures such as relative entropy (also known as cross entropy or Kullback–Leibler distance; it is not an actual distance measure but it is easy to construct a distance measure based on it) and mutual information [BBC$^+$00, TK03]. Other possible measures include the normalized version of mutual information [RS93, RSS94], the information score based on Shannon entropy [KB91], and the information potential based on Renyi's entropy [GP02]. Entropy can be seen as a natural way of measuring the uncertainty, or the information content, of an event using probability distributions. It can also be used to evaluate the difficulty of a decision problem, and different indices based on entropy can also be useful in measuring a classifier performance.

The intra-cluster and inter-cluster densities can also be used as a basis for clustering evaluation [OZ00, WC04]. Specific indices such as Davies–Bouldin, Dunn, and Calinski–Harabasz [BP98, BM01, MB02] have also been generated. These all are based on within-cluster scatter and between-cluster separation, or the diameters of the clusters, and they intrinsically fix a distance measure. If there are hyperspherical or structural clusters, indices based on graph theory [PB97] could be used: a graph structure (minimum spanning tree, relative neighborhood graph, or Gabriel graph) is induced on the partitioned dataset, after which cluster validity indices can be defined. These indices can also be used in defining the correct number of clusters.

The silhouette width criterion [Rou87, VBJ$^+$00] is based on average dissimilarities of the data points of a cluster and the average distances between points inside different clusters. The silhouette width is defined in such a way that its possible values lie between -1 and 1. An object is be said to be well

classified if its silhouette width is 1, and badly classified if the value is negative, since then it is, on average, closer to objects in another cluster. This criterion can also be used for selecting the number of clusters.

Design principles for defining cluster validity indices have also been presented along with new indices [KR05]. As the authors point out, the design principles based on compactness and separability of the clusters have seldom been clearly suggested, and thus some existing indices have features contradicting the main idea of cluster validity indices.

Different validation methods with different clustering methods have been studied by several researchers [MI80, Mil81, MSS83, MC86, MB02]. The problem of replicating cluster analysis has been studied by Breckenridge [Bre89, Bre00].

Chapter 3

On networks

A graph is a collection of vertices and edges connecting them. The edges can be directed or undirected, and real-valued weights may be assigned with the edges. The degree of a vertex is the number of edges connected to the vertex. A path is a sequence of vertices connected together with edges; if the first and the last vertex of the path are the same, the path is a cycle. If every vertex can be reached from every other vertex, the graph is connected. A tree is a connected, acyclic graph; if the number of vertices is n, there are $n-1$ edges in the tree, and removing an edge leads to two disconnected subtrees. A network means a weighted graph, either directed or undirected. A minimum spanning tree (MST) of an undirected weighted graph is a tree containing all the vertices of the network, and the edges are selected in such a way that the total sum of edge weights is smallest possible. [Tar86]

Graph theory is usually said to be founded by Euler, who introduced a solution to the famous Königsberg bridge problem. At first, the studied network problems were small, containing at most a few hundred vertices. Nowadays, the availability of computers allows the collation and analysis of more data, and the focus in network study has shifted to the examination of large-scale statistical properties of the networks. The study of real-world networks, such as social, information, technological, and biological, has revealed that they usually are complex, irregular, or dynamically evolving. The simple network models are not diverse enough to be able to model these real-world networks, and thus complex network models have emerged.

3.1 Complex network models

The complex network models most often studied in the literature include random, small-world and scale-free network models. In the simplest random network model, vertex pairs are connected with each other with some probability value. There exist other random network models, such as a model producing network with power law degree distribution [ACL00].

A small-world network can be constructed starting from a regular lattice, in which each vertex is joined to its neighbors k or fewer lattice spacings away, and then adding or moving a portion of the edges. The moving can be done by examining each vertex in turn and with some probability moving the other end of the edge to a different vertex chosen at random. The result is a lattice with "shortcuts", i.e., the distance between any two vertices is short. Small-world network model was created to lie between regular and random networks [WS98], and the definition has been refined in order to be better suited in the study of real-world networks [LM02]. The small-world phenomenon is also well-known from social networks of human relationships: any two people in the world are most likely linked by a short chain of acquaintances [HA04].

To study the degree distribution of a network, let p_k denote the fraction of vertices with degree k, or, the probability that a vertex chosen at random has degree k. Random network models produce usually a Poisson distribution for p_k, whereas most real-world networks have highly right-skewed degree distributions, meaning that there are lots of vertices having a few connections, and some vertices have many connections — highly-connected vertices are practically absent in random and small-world networks. The networks with right-skewed degree distributions have no characteristic scales for the degrees, hence the networks of this kind are called scale-free. Their degree distribution follows a power law $p_k \sim k^{-\gamma}$. The exponent in the power law has been approximated for many different real-world networks, and it usually has values in range $2 < -\gamma < 4$. [AB02, DM02, New03, Str01]

In order to represent numerically the structural properties of networks, some measures have been defined. These include characteristic path length and clustering coefficient [WS98]. Characteristic path length is a global property measuring the typical path length between two vertices in the network. Clustering coefficient is a local property measuring the connectedness of the neighbors of the vertex. The clustering coefficient of a vertex can be calculated as the ratio of the number of edges between the neighbors of the vertex and the number of all possible edges between the neighbors. The characteristic clustering coefficient of the network can be taken to be the algebraic average of the

clustering coefficients of all the vertices. The clustering coefficients of many real-world networks have been found to be larger than in random networks, which tells that their structure is small-world. [SAK02a] Clustering coefficients have been used in, for example, DNA sequence analysis [GLC06]. It is also possible to define higher-order clustering coefficients [FHJS02].

It is also possible to calculate the values of information-theoretic features, such as the mutual information, the noise level (using conditional entropy), and joint entropy, from networks [SV04]. A measure called neural complexity, based on entropy, has been introduced as a possible way of measuring brain complexity [TSE94]. The study of mammalian brains using graph-theoretic and information-theoretic approaches seems to promise useful results [SE04]. The study of biological networks, such as protein-protein interaction maps, is also a growing research area [BOW04]. The eigenvalue spectra of networks have also been found a useful tool [DGMS04, FDJ$^+$02]. There has also been research concerning evolving networks [JK01, LHN05].

3.2 Properties of scale-free networks

The simplest model for creating scale-free networks is the Barabási–Albert model [BA99, BAJ99]. In contrast to random and small-world models, where the number of vertices is fixed to n, the network is constantly growing by the addition of new vertices, and they attach preferentially to vertices which are already well-connected. The power law degree distribution rises from preferential attachment; if the process is disturbed, scale-free structure does not emerge. There are two possible situations. In the first situation, the vertices are aging in such a way that after a certain amount of time they cannot get any more connections. The second situation happens when the capacity of vertices is limited, meaning that the number of connections is bounded. These networks can still be small-world, but they are broad-scale or single-scale instead of scale-free. [ASBS00] The shape of the attachment probability function, as a function of vertex degree, defines the structure of the network: if the function is asymptotically linear, the resulting structure is scale-free; if the function grows faster than linear, a single vertex connected to nearly every other vertex emerges; if the growth is slower than linear, the number of vertices with degree k decays faster than a power law [KR01].

Scale-free networks can also arise from the situation where there is a fitness-dependent growth rate, meaning that high connectivity and high fitness value of a vertex are preferred [ER02]. In fact, it has been shown that the fitness values

themselves are enough to create a scale-free network, if the connections are made with a probability depending on the fitness values [CCDLRM02, MMK04]. In that model, a vertex with a high fitness value is more highly connected than a vertex with a low fitness value.

The ”rich get richer” -behavior of scale-free networks can be seen, for example, in World Wide Web. However, there are some page subcategories in which the link distribution is different. With this idea in mind, a new network model was introduced to allow the new, poor vertices to gain connections and compete with the older, more connected vertices [PFL$^+$02].

Other complex network models producing power law degree distribution include a simple deterministic model where no probabilities are needed [BRV01], a method for constructing scale-free trees and arbitrary scale-free random networks [BCK01], a model based on the stochastic mean-field model of distance [Ald04], a hybrid model consisting of a global and a local network [CL04], using random walkers [SK04], aggregation (merging vertices together) [AD05], and simple rules creating a scale-free network with adjustable degree distribution exponent [Sta06].

The average distance between the vertices in a scale-free network depends logarithmically on the number of vertices, whereas the probability distribution function of the distances may take different forms [SAK02b]. It has also been shown that if the edge weights are random, then the minimum spanning trees of scale-free networks are scale-free as well [SAK03]. The claim that scale-free networks are self-similar, or that they consist of self-repeating patterns on all length scales, has been verified for a variety of real-world networks [SHM05].

A measure in a scale-free network representing a data packet transport problem, load of a vertex, or the accumulated sum of a fraction of data packets travelling along the shortest pathways between every pair of vertices, has been introduced. It has been shown that the load distributions for many real-world networks follow a power law with an exponent close to 2.2 or 2.0 [GOJ$^+$02, GOG$^+$04].

Scale-free networks are known to have good error tolerance, meaning that if a random vertex is deactivated, the probability that the functionality of the network is hindered is small. On the other hand, they are vulnerable to intentional attacks: if the attacker knows which vertices in the network are the most highly-connected and deactivates them, the network breaks down into small fragments. The attacks can also be aimed at the connections. [AJB00, KCbAH04, LMN04, WC02]

Chapter 4

On graph-theoretic clustering

The concepts of graph theory makes it suitable to be used in clustering problems: the vertices of a weighted graph correspond with the data points in feature space, and the edge weights are the dissimilarities between the pairs of data points [XWI05]. Clustering algorithms based on graph theory are able to find clusters of various shapes, if they are well separated [TK03]; as it was stated in section 2.1, variously shaped clusters are usually problematic to handle.

4.1 MST-based clustering

The most well-known graph-theoretic clustering method is based on the minimum spanning tree (MST) problem [GH85]. The method is very simple: first, an MST of the dataset is constructed, then all the inconsistent edges are searched and deleted [Zah71]. The definition of ”inconsistent” is the most problematic part of the method, and different definitions have been proposed as well as criticized [TK03, DGS83, DH86]. Single-link clustering method, which is a hierarchical method, can also be represented with an MST as follows [JD88].

- Find an MST of the graph. Set the initial clustering: each object belongs in its own cluster.
- Repeat the following two steps until all objects are in one cluster:
  - Find the smallest-weighed edge of the MST. Merge the corresponding clusters.

– Replace the weight of the edge selected in the previous step by a weight larger than the largest distance (i.e., mark the edge as "used").

As it can be seen from the algorithm, the information for single-link clustering is completely contained in the MST [GR69].

There are many algorithms available for MST construction. Three inherently different approaches to the MST problem give rise to three algorithms most commonly used in the literature. They are usually named after the persons who proposed them (actually, many different people were working with the same problems independently, and the names are not necessary the original inventors' names). All the algorithms have as the input a weighted, undirected graph, with n vertices and m edges, and each returns an MST of the graph.

Kruskal The algorithm constructs a forest of trees by adding edges to it in increasing weight order, in such a way that no cycles are formed, and during the construction the trees of the forest are eventually merged together to the MST.

Prim Prim's algorithm starts from an arbitrary vertex (or from the shortest edge). At each step, a shortest possible edge expanding the current MST is added.

Borůvka This algorithm is inherently parallel: the shortest edge for every tree in the current forest is added at the same time. If there are edges with the same weight, some modifications are needed.

All three algorithms can be implemented to run in time $O(m \log n)$. In the case of Prim's algorithm, this requires that the edges are kept in a binary heap; the theoretical time complexity can be further improved by using Fibonacci heaps [GGST86]. Interestingly, it has been shown that the invasion percolation model in physics is essentially the same thing as Prim's algorithm [Bar96]. [HJS84, Tar86, GH85, CLRS01]

The time complexity of deterministic, comparison-based MST construction has been steadily improved [Cha97, Cha00a, PR02b], and as a by-product, a data structure for an approximate priority queue, the soft heap, has been presented [Cha00b]. If better time complexity is wanted, it is possible to find MSTs using a deterministic algorithm not based on comparison [FW94], a randomized algorithm [KKT95, PR02a], a parallel algorithm [CHL01], or a parallel randomized algorithm [PR99].

Another two viewpoints to the MST problem are the MST verification problem, or determining whether a given spanning tree is an MST or not [Tar79,

Kin97, BKRW98], and the maintaining of an MST of a dynamic graph [Epp95, ACI97, HK97, Zar02, CFFPI02].

4.2 Other graph-theoretic clustering methods

A graph-theoretic approach to clustering using undirected adjacency graphs has been introduced by Wu and Leahy. [WL93]. The clustering is achieved by removing edges to form subgraphs in such a way that the largest inter-subgraph maximum flow is minimized. The clustering algorithm has been applied to image segmentation problem, and it is said to be able to accurately locate region boundaries in the images.

A hierarchical clustering method called diameter clustering, based on an Euclidean MST, has been presented by Krznaric and Levcopoulos [KL95]. The diameter clusters are defined in such a way that they are well-separated and form a hierarchy which can be described by a tree. This structure can be used in some other applications such as in calculating the complete link hierarchy.

A generalized distance measure in undirected weighted graphs suitable to clustering has been introduced [Len98]. It has been used in a two-step clustering procedure, where the cluster centers are found with either crisp or fuzzy k-means, and the samples near the borders of the clusters are handled with a graph-theoretic method similar to single-linkage method.

A graph-theoretic clustering algorithm using a similarity graph in which clusters are highly connected subgraphs has also been presented [HS00]. The subgraphs are found using minimum cut algorithms; there is no need to know the number of clusters. The method has been tested with gene expression data, and it has been found to be efficient.

An another graph-theoretic clustering algorithm using minimum weight cuts has been introduced [SS00]. The data to be clustered is represented by a similarity matrix, and in the graph, there is an edge between two nodes, if their similarity measure is above some pre-defined non-negative threshold. The method, called CLICK, has been tested on biological datasets. It is also said to be a very fast method.

A pairwise nearest neighbor (PNN) clustering method using k-nearest neighbor graph has been presented by Fränti et al. [FVH03]. In this method, k-nearest neighbor graph represents a cluster for each vertex, and edges are pointers to neighboring clusters. The PNN method is a hierarchical clustering method, in which clusters are merged together, and the graph has been used to assist in that purpose.

Three graph clustering algorithms, namely Markov clustering, iterative conductance cutting, and geometric MST, have been presented along with the comparison of their performance in the case of randomly generated datasets [BGW03]. All the methods are meant to separate sparsely connected dense subgraphs from each other using indices measuring intra-cluster density and inter-cluster sparsity.

An MST clustering method automatically detecting inconsistent edges has been proposed [HC03]. The algorithm finds statistically the threshold for the edges to be removed, and connects very small clusters to the nearest cluster. The method has been tested with artificial datasets.

Three methods for obtaining a clustering from an MST, including the removal of long edges, partitioning the MST iteratively, and finding the globally optimal clustering for a range of number of clusters, have been proposed [XOX01, XOX02]. The methods have been tested with gene expression data. It is also possible to formulate an exact definition for a cluster in this framework [OXX03].

A shape-independent clustering technique based on iterative partitioning of the relative neighborhood graph has been given by Bandyopadhyay [Ban04]. The proposed method is claimed to be able to detect outliers, indicate the inherent hierarchical nature of the clusters present in the dataset, and identify the case when there are no natural clusters in the data. The method has been tested with artificial datasets only.

Ordering the edges of an MST has also been proposed as a clustering method [FCOCC04]. This method has been tested with simulated and real datasets, and the results were compared with the results of a similar clustering method which orders the data points using the local density.

Recently, a nonparametric clustering algorithm being able to find clusters of varying shapes and overlapping clusters has been presented [ZZZL07]. The method is in close relationship with the single-link clustering method. Its basis is nonparametric density estimation. Thus, there is no restriction for the shape of the density function.

A partitional clustering algorithm based on graph coloring with a greedy algorithm being able to select the appropriate number of clusters based on a clustering tendency index has been proposed by Brás Silva [BSBPdC06]. The testing has shown the efficiency of the method, also in the case when there is no clustering structure in the dataset.

4.3 Applications of graph-theoretic clustering

The problem with image database querying has been solved with graph-theoretic clustering algorithm [AH99]. The clustering is used as a post-processing step after the database has been queried: the best matching images are queried again, after which the images are arranged into a graph whose connected clusters include the best matching images. The clusters can overlap, which is a desired property.

A memetic clustering algorithm, or a genetic algorithm combined with local search, for clustering gene expression data has been introduced [SMSZ03]. Before the actual memetic algorithm is used, the MST of the dataset is calculated. The initial population is then set to be the MST from which $k-1$ random edges have been deleted. This combination is said to find near-optimal solutions quickly.

A method for automatically finding cluster of micro-calcifications from mammographic images has been given by Cordella et al. [CPSV05]. In this application area, a cluster is a group of at least three micro-calcifications in a limited area of the mammogram. The clustering method begins by assigning the vertices with the micro-calcifications found by a detection algorithm and the edges with their Euclidean distances. Then the MST of this graph is determined, and the edges with large weights are removed with the help of a fuzzy c-means algorithm.

A graph-theoretic approach can also be used in image segmentation problems [MK95]. Since the pixel data contained in an image is vast, the cluster analysis is done using the histogram of the image. A directed graph is formed by calculating the number of pixels in each histogram bin, and drawing a link from each bin to the bin with maximum count in its neighborhood. The directed, rooted trees are the clusters found from the image. The method has been used in a binarization algorithm for color text images, and it has been found to be effective [WLLH05].

An another solution for image segmentation problem, sparse clustering, has been provided by Jeon et al. [JJH06]. The image is presented with a weighted undirected graph, and the segmentation is acquired by factorizing the affinity matrix using positive tensor factorization. The number of clusters is detected automatically with the help of intra- and inter-cluster measures.

Chapter 5

Summary of publications and results

The first publication, ”Minimum spanning tree clustering of EEG signals”, **I**, studies the usage of the standard minimum spanning tree in clustering of EEG signals containing epileptic seizures. Three strategies to find inconsistent edges from the MST were presented, each based on the statistics of edge lengths. The clustering results obtained were compared with the results from k-means. The problems associated with the MST clustering method became apparent, although it was also found out that the method seemed to place similar-looking fragments of EEG signal near each other in the resulting tree. In conclusion, it can be said that the EEG dataset used in the study was difficult to cluster with k-means also, and the use of the MST method can be justified since clusters of different size were preferred in this application area. It can also be noted that k-means does not reveal possible outlier values whereas the MST method can find them (and thus create singleton clusters).

In ”Clustering with a minimum spanning tree of scale-free-like structure”, **II**, the scale-free minimum spanning tree (SFMST) structure was first presented along with its usage in clustering. The performance of SFMST clustering method was compared with MST clustering and k-means. Test datasets included three freely available datasets as well as a dataset containing EEG recordings with epileptic seizures. All datasets consisted of continuous-valued features. This article forms the basis of the thesis.

The main idea of the algorithm presented in **II** is as follows. First, the distance matrix (containing the distances between all the data points) is calculated. The initial edge weights are set to the reversed distances, that is,

$w_0(i,j) = \lceil \max_{i,j}(d(i,j)) \rceil - d(i,j)$, where $d(i,j)$ is the Euclidean distance between vertices i and j, and the ceiling operation $\lceil \cdot \rceil$ is used to avoid any edge having weight zero. Hence, in this formulation, the edge with the greatest weight corresponds with the shortest distance. During the execution of the algorithm, the SFMST is constructed by selecting the edge with the greatest weight, checking if the edge can be added to the growing SFMST (in a way that no cycles are formed), and updating the weights of the edges not in the SFMST, if necessary. The necessity is defined using a pre-defined threshold value: if a vertex has more edges than the threshold, the weights of all the edges having the vertex as one endpoint have to be updated. The updating function was defined as $w_{\mathrm{new}} = w_0 + n\,c^n$, where n is the number of the edges, and c is a constant whose possible values are $0.5 < c < 1$. The reason behind the selection of the weight updating function was that the bonus for high connectivity increases slowly when the number of connections increases, and starts to decrease when the number of connections is large enough. The two constants, the threshold value and c, define the structure of the resulting SFMST. The role of c is especially important: if $c = 1$, the resulting SFMST has one vertex to which every other vertex is directly connected to. Setting a smaller value for c causes the emergence of more hubs, or highly-connected vertices. The smaller value of c leads to smaller hubs having less connections, and at $c = 0.5$, the resulting SFMST looks nearly like the standard MST. The constant c might be called ”a hubability constant”.

The performance of SFMST clustering method was found to strongly depend on the value of c: the same value for c was used with all the datasets, although the optimal result would require the selection of the constant according to the characteristics of the dataset. It is also possible to use different methods in finding the clustering from the SFMST. A vertex was defined as a hub if it had at least four connections, and a cluster was defined to consist of the hub and all the vertices connecting directly to it. Two hubs connected to each other either directly or via one linking vertex were defined to belong to the same cluster. In addition, the SFMST structure may have branches, or chains of vertices originating from a hub.

The results can easily be compared with the ones from MST clustering method, whereas k-means method is inherently different method which makes comparison harder. The SFMST outperformed the standard MST method, but k-means performed better than SFMST method with certain datasets. The MST clustering method was found to produce lots of small clusters with only a few members, both in **I** and **II**. These small clusters could be interpreted to

contain outlier points. The SFMST method produces branches whose role is unclear — they might also be interpreted as outliers. In addition, there may be many hubs inside one cluster.

The SFMST clustering method was improved in "Modifying the scale-free clustering method", **III**. In this article, a new method for edge weight updating was presented: the edge weights are always updated after an edge is added to the SFMST. The edge weight updating function was also changed to resemble the equation for the gravitational force between two particles. In more formal way, the initial edge weights were defined as $w_0(i,j) = 1/d(i,j)^2$. The weight updating function became $w_{\mathrm{new}}(i,j) = nc^n/d(i,j)^2$, where, as before, $0.5 < c < 1$, and n is the number of edges. The improvement made the SFMST method simpler: now it needs only one parameter, the hubability constant c.

The improved method was tested with three freely available datasets having features with continuous values, and the results were compared with the results from k-means method. The improved SFMST method produced better results than k-means in the case of the first dataset; with the second datasets, the performances were about equal quality. The third dataset was not well separated with either of the methods.

"Finding the optimal number of clusters from artificial datasets", **IV**, presents a way to find the number of clusters as well as the clustering itself from the SFMST structure. The modified SFMST construction method was used along with a new way of defining the edges to be removed. First, a histogram of the edge lengths in the SFMST was constructed (the number of bins needed was detected automatically); it was found out that the edge length distribution was best modeled using a lognormal distribution. The histogram can be truncated at the point where it first reaches zero, the edges corresponding with the isolated bins in the histogram removed. As a test data, three artificial datasets were used. The results of the SFMST method were compared with the results from nearest neighbor and k-means clustering, for which the number of clusters was detected using the largest average silhouette width criterion. Nearest neighbor method was used since it is closely connected to the MST clustering method.

The results showed that the SFMST method tended to find greater number of clusters than were actually present in the data. The clusterings itselves were of quite a good quality with respect to a priori information about the datasets. The same value for c was used in all cases; selecting c individually for each dataset might have lead to better clustering results. The nearest neighbor method was found to perform better than the k-means method with the artificial datasets. This follows probably from the fact that the datasets were designed

to have clusters with different shapes and densities.

The last publication, ”A fast clustering method using a scale-free minimum spanning tree”, **V**, focuses on the computational complexity of the SFMST. It was shown that an SFMST can be efficiently constructed when binary heaps are used. One of the goals was also to check the practical computational complexity of Fibonacci heaps: it was known beforehand that their low asymptotic time complexity requires large constant factors, which means that their maintenance takes a lot of time in practice; this was confirmed in the study.

Chapter 6

Conclusions and discussion

In this study, the clustering problem found in many different application areas was discussed from the viewpoint of networks. A novel graph-theoretic clustering method using a structure called scale-free minimum spanning tree was formulated. The method is capable of handling data with continuous-valued attributes and it needs only one control parameter. The number of clusters present can be detected during the clustering process. The method is able to find clusters with different sizes, shapes and densities, and it has been tested with both real and artificial datasets. To the best of our knowledge, this is the first time anyone has used a scale-free tree as a way to obtain a clustering.

Open questions related to the proposed method include the selection of an appropriate dissimilarity measure which also defines the metrics to be used in validating, and automatically finding the most suitable value of the hubability constant c for each dataset in the SFMST construction. The first problem is universal in clustering; the similarity measure has to be chosen separately for each dataset. The second one, if it proves to be solvable, makes the method fully automatic. To make the method more effective computationally, randomization could be implemented. More thorough theoretical analysis of the method is also needed. The proposed method might be somehow related to clustering methods based on densities; the possible relation would be useful to know.

The clustering method introduced in this study presents two ways of obtaining the clustering from the SFMST structure, directly from the structure and by using the edge length histogram. There may exist different, better strategies for finding the clusters, depending on how the hubs, branches and edge lengths of the structure are interpreted. Defining the optimal number of clusters can also be included in the strategy.

The significance of this study can be found in the combination of a scale-

free structure and clustering. There is still room for new clustering methods in data mining applications, and importing concepts from graph theory into clustering may reveal hidden structures in the data not found before with other methods. However, care must be taken when using a clustering method: it makes no sense to impose a partition into a dataset which is not inherently partitionable. Therefore, clustering validity has to be taken into account, and the clustering results must be viewed with a critical eye.

Bibliography

[AB02] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[ABKS99] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data,* May 31 - June 3, Philadelphia, Pennsylvania, USA, pages 49–60, New York, USA, 1999. ACM Press.

[ACI97] D. Alberts, G. Cattaneo, and G. F. Italiano. An empirical study of dynamic graph algorithms. *Journal of Experimental Algorithmics*, 2:Article No. 5, 1997.

[ACL00] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of the 32nd annual ACM symposium on theory of computing (STOC'00),* May 21–23, Portland, Oregon, USA, pages 171–180, New York, USA, 2000. ACM Press.

[AD05] M. J. Alava and S. N. Dorogovtsev. Complex networks created by aggregation. *Physical Review E*, 71(3):36107, 2005.

[AH99] S. Aksoy and R. M. Haralick. Graph-theoretic clustering for image grouping and retrieval. In *Proceedings of 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 23-25 June, Fort Collins, Colorado, USA, volume 1, pages 63–68, Los Alamitos, California, USA, 1999. IEEE.

[AJB00] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[Ald04] D. J. Aldous. A tractable complex network model based on the stochastic mean-field model of distance. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 51–87, 2004.

[AMN+98] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[AP02] P. K. Agarwal and C. M. Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002.

[ASBS00] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149–11152, 2000.

[BA99] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[BAJ99] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272(1–2):173–187, 1999.

[Ban04] S. Bandyopadhyay. An automatic shape independent clustering technique. *Pattern Recognition*, 37(1):33–45, 2004.

[Bar96] A.-L. Barabási. Invasion percolation and global optimization. *Physical Review Letters*, 76(20):3750–3753, 1996.

[BBC+00] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[BBM06] P. Bertrand and G. Bel Mufti. Loevinger's measures of rule quality for assessing cluster stability. *Computational Statistics & Data Analysis*, 50(4):992–1015, 2006.

[BCK01] Z. Burda, J. D. Correia, and A. Krzywicki. Statistical ensemble of scale-free random graphs. *Physical Review E*, 64(4):46118, 2001.

[BGRS99] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is ”nearest neighbor” meaningful? In *Proceedings of the 7th International Conference on Database Theory,* 10-12 January, Jerusalem, Israel, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235, Berlin Heidelberg, 1999. Springer Verlag.

[BGW03] U. Brandes, M. Gaertler, and D. Wagner. Experiments on graph clustering algorithms. In *Proceedings of the 11th European Symposium on Algorithms (ESA'03),* 15-20 September, Budapest, Hungary, volume 2832 of *Lecture Notes in Computer Science*, pages 568–579, Berlin Heidelberg, 2003. Springer Verlag.

[Bic03] D. R. Bickel. Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically. *Bioinformatics*, 19(7):818–824, 2003.

[BKRW98] A. L. Buchsbaum, H. Kaplan, A. Rogers, and J. R. Westbrook. Linear-time pointer-machine algorithms for least common ancestors, MST verification, and dominators. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing,* 24-26 May, Dallas, Texas, USA, pages 279–288, New York, USA, 1998. ACM.

[BM00] V. Batagelj and A. Mrvar. Some analyses of Erdös collaboration graph. *Social Networks*, 22(2):173–186, 2000.

[BM01] S. Bandyopadhyay and U. Maulik. Nonparametric genetic clustering: comparison of validity indices. *IEEE Transactions on Systems, Man and Cybernetics. Part C: Applications and Reviews*, 31(1):120–125, 2001.

[BM02] S. Bandyopadhyay and U. Maulik. Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognition*, 35(6):1197–1208, 2002.

[BOW04] A.-L. Barabási, Z. N. Oltvai, and S. Wuchty. Characteristics of biological networks. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 443–457, 2004.

[BP98] J. C. Bezdek and N. R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics*, 28(3):301–315, 1998.

[BR93a] S. Banerjee and A. Rosenfeld. Model-based cluster analysis. *Pattern Recognition*, 26(6):963–974, 1993.

[BR93b] J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[Bre89] J. N. Breckenridge. Replicating cluster analysis: method, consistency, and validity. *Multivariate Behavioral Research*, 24(2):147–161, 1989.

[Bre00] J. N. Breckenridge. Validating cluster analysis: consistent replication and symmetry. *Multivariate Behavioral Research*, 35(2):261–285, 2000.

[BRV01] A.-L. Barabási, E. Ravasz, and T. Vicsek. Deterministic scale-free networks. *Physica A*, 299(3–4):559–564, 2001.

[BSBPdC06] H. Brás Silva, P. Brito, and J. Pinto da Costa. A partitional clustering algorithm validated by a clustering tendency index based on graph theory. *Pattern Recognition*, 39(5):776–788, 2006.

[CCDLRM02] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Munoz. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89(25):258702, 2002.

[CCP06] J. Caiado, N. Crato, and D. Peña. A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, 50(10):2668–2684, 2006.

[CDW06] S. S. Chae, J. L. DuBien, and W. D. Warde. A method of predicting the number of clusters using Rand's statistic. *Computational Statistics & Data Analysis*, 50(12):3531–3546, 2006.

[CFFPI02] G. Cattaneo, P. Faruolo, U. Ferraro Petrillo, and G. F. Italiano. Maintaining dynamic minimum spanning trees: an experimental study. In *Proceedings of the 4th International Workshop on Algorithm Engineering and Experiments (ALENEX 2002),* January 4-5, San Francisco, California, USA, volume 2409 of *Lecture Notes in Computer Science*, pages 111–125, Berlin Heidelberg, 2002. Springer Verlag.

[CG95] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[Cha97] B. Chazelle. A faster deterministic algorithm for minimum spanning trees. In *Proceedings of the 38th IEEE Annual Symposium on Foundations of Computer Science,* 20–22 October, Miami Beach, Florida, USA, pages 22–31, Los Alamitos, California, USA, 1997. IEEE.

[Cha00a] B. Chazelle. A minimum spanning tree algorithm with inverse-Ackermann type complexity. *Journal of the ACM*, 47(6):1028–1047, 2000.

[Cha00b] B. Chazelle. The soft heap: An approximate priority queue with optimal error rate. *Journal of the ACM*, 47(6):1012–1027, 2000.

[CHL01] K. W. Chong, Y. Han, and T. W. Lam. Concurrent threads and optimal parallel minimum spanning trees algorithm. *Journal of the ACM*, 48(2):297–323, 2001.

[CL04] F. Chung and L. Lu. The small world phenomenon in hybrid power law graphs. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 89–104, 2004.

[CLRS01] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. McGraw-Hill, Boston, second edition, 2001.

[CNBYM01] E. Chavez, G. Navarro, R. Baeza-Yates, and J. L. Marroquin. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

[Coh02] D. Cohen. All the world's a net. *New Scientist*, 174(2238):24–29, 2002.

[CPSV05] L. P. Cordella, G. Percannella, C. Sansone, and M. Vento. A graph-theoretical clustering method for detecting clusters of micro-calcifications in mammographic images. In *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems,* 23-24 June, Dublin, Ireland, pages 15–20, Los Alamitos, California, USA, 2005. IEEE.

[DDW02] E. Dimitriadou, S. Dolničar, and A. Weingessel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1):137–159, 2002.

[DF02] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7):research0036, 2002.

[DGMS04] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, and A. N. Samukhin. Spectral analysis of random networks. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 35–50, 2004.

[DGS83] V. Di Gesu and B. Sacco. Some statistical properties of the minimum spanning forest. *Pattern Recognition*, 16(5):525–531, 1983.

[DH86] R. C. Dubes and R. L. Hoffman. Remarks on some statistical properties of the minimum spanning forest. *Pattern Recognition*, 19(1):49–53, 1986.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, Inc., New York, second edition, 2001.

[DM02] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002.

[DR98] A. Dasgupta and A. E. Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302, 1998.

[EC02] V. Estivill-Castro. Why so many clustering algorithms. *ACM SIGKDD Explorations*, 4(1):65–75, 2002.

[Epp95] D. Eppstein. Dynamic Euclidean minimum spanning trees and extrema of binary functions. *Discrete & Computational Geometry*, 13:111–122, 1995.

[ER02] G. Ergün and G. J. Rodgers. Growing random networks with fitness. *Physica A*, 303(1–2):261–272, 2002.

[ESK03] L. Ertöz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of SIAM International Conference on Data*

Mining 2003, May 1-3, San Francisco, CA, USA, Philadelphia, Pennsylvania, USA, 2003. SIAM.

[FCOCC04] M. Forina, M. C. Cerrato Oliveros, C. Casolino, and M. Casale. Minimum spanning tree: ordering edges to identify clustering structure. *Analytica Chimica Acta*, 515(1):43–53, 2004.

[FDJ+02] I. Farkas, I. Derényi, H. Jeong, Z. Néda, Z. N. Oltvai, E. Ravasz, A. Schubert, A.-L. Barabási, and T. Vicsek. Networks in life: Scaling properties and eigenvalue spectra. *Physica A*, 314(1–4):25–34, 2002.

[FHJS02] A. Fronczak, J. A. Hołyst, M. Jedynak, and J. Sienkiewicz. Higher order clustering coefficients in Barabási–Albert networks. *Physica A*, 316(1–4):688–694, 2002.

[Fin05] H. Finch. Comparison of distance measures in cluster analysis with dichotomous data. *Journal of Data Science*, 3(1):85–100, 2005.

[FIP98] J. W. Fisher III and J. C. Principe. A methodology for information theoretic feature extraction. In *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks (IJCNN'98),* May 4-9, volume 3, pages 1712–1716, Los Alamitos, California, USA, 1998. IEEE.

[FJ02a] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[FJ02b] A. L. N. Fred and A. K. Jain. Data clustering using evidence accumulation. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02),* Quebec, Canada, volume 4, pages 276–280, 2002.

[FK96] H. Frigui and R. Krishnapuram. A robust algorithm for automatic extraction of an unknown number of clusters from noisy data. *Pattern Recognition Letters*, 17(12):1223–1322, 1996.

[FL03] A. L. N. Fred and J. M. N. Leitão. A new cluster isolation criterion based on dissimilarity increments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):944–958, 2003.

[FR98] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[FU02] U. Fayyad and R. Uthurusamy. Evolving data mining into solutions for insights. *Communications of the ACM*, 45(8):28–31, 2002.

[Fuk88] K. Fukushima. A neural network for visual pattern recognition. *Computer*, 21(3):65–75, 1988.

[FVH03] P. Fränti, O. Virmajoki, and V. Hautamäki. Fast PNN-based clustering using k-nearest neighbor graph. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),* November 19-22, Melbourne, Florida, USA, pages 525–528, 2003.

[FW94] M. L. Fredman and D. E. Willard. Trans-dichotomous algorithms for minimum spanning trees and shortest paths. *Journal of Computer and System Sciences*, 48(3):533–551, 1994.

[GCL02] P. Guo, C. L. P. Chen, and M. R. Lyu. Cluster number selection for a small set of samples using the Bayesian Ying-Yang model. *IEEE Transactions on Neural Networks*, 13(3):757–763, 2002.

[GGST86] H. N. Gabow, Z. Galil, T. Spencer, and R. E. Tarjan. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, 6(2):109–122, 1986.

[GH85] R. L. Graham and P. Hell. On the history of the minimum spanning tree problem. *Annals of the History of Computing*, 7(1):43–57, 1985.

[GLC06] G. J. L. Gerhardt, N. Lemke, and G. Corso. Network clustering coefficient approach to dna sequence analysis. *Chaos, Solitons and Fractals*, 28(4):1037–1045, 2006.

[GOG+04] K.-I. Goh, E. Oh, C.-M. Ghim, B. Kahng, and D. Kim. Classes of the shortest pathway structures in scale free networks. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 105–125, 2004.

[GOJ+02] K.-I. Goh, Eu. Oh, H. Jeong, B. Kahng, and D. Kim. Classification of scale-free networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12583–12588, 2002.

[GP02] E. Gokcay and J. C. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):158–171, 2002.

[GR69] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18(1):54–64, 1969.

[Gro05] J. W. Grossman. Patterns of research in mathematics. *Notices of the AMS*, 52(1):35–41, 2005.

[HA04] B. A. Huberman and L. A. Adamic. Information dynamics in the networked world. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 371–398, 2004.

[Hak05] H. Haken. Synchronization and pattern recognition in a pulse-coupled neural net. *Physica D*, 205(1–4):1–6, 2005.

[Har96] A. Hardy. On the number of clusters. *Computational Statistics & Data Analysis*, 23(1):83–96, 1996.

[Hay94] S. Haykin. *Neural networks. A comprehensive foundation*. Prentice Hall International, Inc., London, 1994.

[HBH05] J. M. Huband, J. C. Bezdek, and R. J. Hathaway. bigVAT: Visual assessment of cluster tendency for large data sets. *Pattern Recognition*, 38(11):1875–1886, 2005.

[HBV96] M. Herbin, N. Bonnet, and P. Vautrot. A clustering method based on the estimation of probability density function and on the skeleton by influence zones. Application to image processing. *Pattern Recognition Letters*, 17(11):1141–1150, 1996.

[HBV01] M. Herbin, N. Bonnet, and P. Vautrot. Estimation of the number of clusters and influence zones. *Pattern Recognition Letters*, 22(14):1557–1568, 2001.

[HBV02a] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster validity methods: Part I. *ACM SIGMOD Record*, 31(2):40–45, 2002.

[HBV02b] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering validity checking methods: Part II. *ACM SIGMOD Record*, 31(3):19–27, 2002.

[HC03] Y. He and L. Chen. A novel nonparametric clustering algorithm for discovering arbitrary shaped clusters. In *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, and the Fourth Pacific Rim Conference on Multimedia (ICICS-PCM 2003),* 15-18 December, Singapore, volume 3, pages 1826–1830, Los Alamitos, California, USA, 2003. IEEE.

[HE03] E. R. Hruschka and N. F. F. Ebecken. A genetic algorithm for cluster analysis. *Intelligent Data Analysis*, 7(1):15–25, 2003.

[HJS84] R. E. Haymond, J. P. Jarvis, and D. R. Shier. Computational methods for minimum spanning tree algorithms. *SIAM Journal on Scientific and Statistical Computing*, 5(1):157–174, 1984.

[HK97] M. R. Henzinger and V. King. Maintaining minimum spanning trees in dynamic graphs. In *Proceedings of Automata, Languages and Programming: 24th International Colloquium (ICALP'97),* July 7-11, Bologna, Italy, volume 1256 of *Lecture Notes in Computer Science*, pages 594–604, Berlin Heidelberg, 1997. Springer Verlag.

[HK01] J. Han and M. Kamber. *Data mining: Concepts and techniques*. Morgan Kaufmann, San Francisco, California, USA, 2001.

[HS00] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4-6):175–181, 2000.

[JD88] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[JDM00] A. K. Jain, R .P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[JJH06] B.-K. Jeon, Y.-B. Jung, and K.-S. Hong. Image segmentation by unsupervised sparse clustering. *Pattern Recognition Letters*, 27(14):1650–1664, 2006.

[JK01] S. Jain and S. Krishna. A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):543–547, 2001.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Fynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[KB91] I. Kononenko and I. Bratko. Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6(1):67–80, 1991.

[KCbAH04] T. Kalisky, R. Cohen, D. ben Avraham, and S. Havlin. Tomography and stability of complex networks. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 3–34, 2004.

[KH04] P. M. Kanade and L. O. Hall. Fuzzy ant clustering by centroid positioning. In *Proceedings of 2004 IEEE International Conference on Fuzzy Systems,* 25-29 July, Budapest, Hungary, volume 1, pages 371–376. IEEE, 2004.

[KHDM98] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[Kin97] V. King. A simpler minimum spanning tree verification algorithm. *Algorithmica*, 18(2):263–270, 1997.

[KKT95] D. R. Karger, P. N. Klein, and R. E. Tarjan. A randomized linear-time algorithm to find minimum spanning trees. *Journal of the ACM*, 42(2):321–328, 1995.

[KL95] D. Krznaric and C. Levcopoulos. Computing hierarchies of clusters from the Euclidean minimum spanning tree in linear time. In *Proceedings of 15th Conference on Foundations of Software Technology and Theoretical Computer Science,* December 18-20, Bangalore, India, volume 1026 of *Lecture Notes in Computer Science*, pages 443–455, Berlin Heidelberg, 1995. Springer Verlag.

[KLV98] S. R. Kulkarni, G. Lugosi, and S. S. Venkatesh. Learning pattern classification—a survey. *IEEE Transactions on Information Theory*, 44(6):2178–2206, 1998.

[Koj04] I. Kojadinovic. Agglomerative hierarchical clustering of continuous variables based on mutual information. *Computational Statistics & Data Analysis*, 46(2):269–294, 2004.

[KP99] R. Kothari and D. Pitts. On finding the number of clusters. *Pattern Recognition Letters*, 20(4):405–416, 1999.

[KR01] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

[KR05] M. Kim and R. S. Ramakrishna. New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005.

[KSP01] S. Kaski, J. Sinkkonen, and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12(4):936–947, 2001.

[Len98] C. Lenart. A generalized distance in graphs and centered partitions. *SIAM Journal of Discrete Mathematics*, 11(2):293–304, 1998.

[LF01] A. V. Lukashin and R. Fuchs. Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, 17(5):405–414, 2001.

[LHN05] E. Lieberman, C. Hauert, and M. A. Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312–316, 2005.

[LM02] V. Latora and M. Marchiori. Is the Boston subway a small-world network? *Physica A*, 314(1–4):109–113, 2002.

[LMN04] Y.-C. Lai, A. E. Motter, and T. Nishikawa. Attacks and cascades in complex networks. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 299–310, 2004.

[MB02] U. Maulik and S. Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

[MC85] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[MC86] G. W. Milligan and M. C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4):441–458, 1986.

[MI80] G. W. Milligan and P. D. Isaac. The validation of four ultrametric clustering algorithms. *Pattern Recognition*, 12(2):41–50, 1980.

[Mil80] G. W. Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3):325–342, 1980.

[Mil81] G. W. Milligan. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199, 1981.

[Mil85] G. W. Milligan. An algorithm for generating artificial test clusters. *Psychometrika*, 50(1):123–127, 1985.

[MK95] J. Matas and J. Kittler. Spatial and feature space clustering: Applications in image analysis. In *Proceedings of the 6th International Conference on Computer Analysis of Images and Patterns (CAIP'95),* September 6-8, Prague, Czech Republic, volume 970 of *Lecture Notes in Computer Science*, pages 162–173, Berlin Heidelberg, 1995. Springer Verlag.

[MMK04] N. Masuda, H. Miwa, and N. Konno. Analysis of scale-free networks based on a threshold graph with intrinsic vertex weights. *Physical Review E*, 70(3):036124, 2004.

[MSS83] G. W. Milligan, S. C. Soon, and L. M. Sokol. The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(1):40–47, 1983.

[New03] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[New04] M. E. J. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. In *Complex Networks*, volume 650 of *Lecture Notes in Physics*, pages 337–370, 2004.

[OXX03] V. Olman, D. Xu, and Y. Xu. Identification of regulatory binding sites using minimum spanning trees. In *Proceedings of the 8th*

Pacific Symposium of Biocomputing (PSB 2003), January 3-7, Hawaii, USA, pages 327–338, 2003.

[OZ00] S. H. Ong and X. Zhao. On post-clustering evaluation and modification. *Pattern Recognition Letters*, 21(5):365–373, 2000.

[PB97] N. R. Pal and J. Biswas. Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6):847–857, 1997.

[PFL+02] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):5207–5211, 2002.

[PLP+05] N. Päivinen, S. Lammi, A. Pitkänen, J. Nissinen, M. Penttonen, and T. Grönfors. Epileptic seizure detection: A nonlinear viewpoint. *Computer Methods and Programs in Biomedicine*, 79(2):151–159, 2005.

[PM00] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000),* June 29 - July 2, pages 727–734, San Francisco, 2000. Morgan Kauffmann.

[PR99] S. Pettie and V. Ramachandran. A randomized time-work optimal parallel algorithm for finding a minimum spanning forest. In *Randomization, Approximation, and Combinatorial Optimization. Algorithms and Techniques: Third International Workshop on Randomization and Approximation Techniques in Computer Science, and Second International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (RANDOM-APPROX'99),* 8-11 August, Berkeley, California, USA, volume 1671 of *Lecture Notes in Computer Science*, pages 233–244, Berlin Heidelberg, 1999. Springer Verlag.

[PR02a] S. Pettie and V. Ramachandran. Minimizing randomness in minimum spanning tree, parallel connectivity, and set maxima algorithms. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'02),* 6-8 January, San Francisco, California, USA, pages 713–722, New York, 2002. ACM.

[PR02b] S. Pettie and V. Ramachandran. An optimal minimum spanning tree algorithm. *Journal of the ACM*, 49(1):16–34, 2002.

[QJ06] W. Qiu and H. Joe. Separation index and partial membership for clustering. *Computational Statistics & Data Analysis*, 50(3):585–603, 2006.

[Rou87] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[RS93] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232(2):584–599, 1993.

[RSS94] B. Rost, C. Sander, and R. Schneider. Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, 235(1):13–26, 1994.

[SAK02a] G. Szabó, M. Alava, and J. Kertész. Clustering in complex networks. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 139–162, 2002.

[SAK02b] G. Szabó, M. Alava, and J. Kertész. Shortest paths and load scaling in scale-free trees. *Physical Review E*, 66(2):26101, 2002.

[SAK03] G. Szabó, M. Alava, and J. Kertész. Geometry of minimum spanning trees on scale-free networks. *Physica A*, 330(1–2):31–36, 2003.

[SE04] A. K. Seth and G. M. Edelman. Theoretical neuroanatomy: Analyzing the structure, dynamics, and function of neuronal networks. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 483–511, 2004.

[SHM05] C. Song, S. Havlin, and H. A. Makse. Self-similarity of complex networks. *Nature*, 433(7024):392–395, 2005.

[SK04] J. Saramäki and K. Kaski. Scale-free networks generated by random walkers. *Physica A*, 341:80–86, 2004.

[SMSZ03] N. Speer, P. Merz, C. Spieth, and A. Zell. Clustering gene expression data with memetic algorithms based on minimum spanning trees. In *The 2003 Congress on Evolutionary Computation (CEC'03),* 8-12 December, Canberra, Australia, volume 3, pages 1848–1855. IEEE, 2003.

[Soh99] S. Y. Sohn. Meta analysis of classification algorithms for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1137–1144, 1999.

[SS00] R. Sharan and R. Shamir. CLICK: A clustering algorithm with applications to gene expression analysis. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-2000),* August 19-23, La Jolla, California, USA, pages 307–316. AAAI Press, 2000.

[Sta06] F. Stauffer. Two-level relationships and scale-free networks. *Physica A*, 365(2):565–570, 2006.

[Str01] S. H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[SV04] R. V. Solé and S. Valverde. Information theory of complex networks: On evolution and architectural constraints. In *Complex networks*, volume 650 of *Lecture Notes in Physics*, pages 189–207, 2004.

[SW99] P. D. Scott and E. Wilkins. Evaluating data mining procedures: techniques for generating artificial data sets. *Information and Software Technology*, 41(9):579–587, 1999.

[Tar79] R. E. Tarjan. Applications of path compression on balanced trees. *Journal of the ACM*, 26(4):690–715, 1979.

[Tar86] R. E. Tarjan. *Data structures and network algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1986.

[TK03] S. Theodoridis and K. Koutroumbas. *Pattern recognition*. Academic Press, Amsterdam, The Netherlands, second edition, 2003.

[TSE94] G. Tononi, O. Sporns, and G. M. Edelman. A measure for brain complexity: Relating functional segregation and integration in

the nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, 91(11):5033–5037, 1994.

[TWH01] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[TY00] L. Y. Tseng and S. B. Yang. A genetic clustering algorithm for data with non-spherical-shape clusters. *Pattern Recognition*, 33(7):1251–1259, 2000.

[VBJ+00] J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen. Mining for putative regulatory elements in the yeast genome using gene expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-2000),* August 19-23, La Jolla, California, USA, pages 384–394. AAAI Press, 2000.

[WC02] X. Fan Wang and G. Chen. Synchronization in scale-free dynamical networks: Robustness and fragility. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(1):54–62, 2002.

[WC04] S. Wu and T. W. S. Chow. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition*, 37(2):175–188, 2004.

[WCS01] C.-C. Wong, C.-C. Chen, and M.-C. Su. A novel algorithm for data clustering. *Pattern Recognition*, 34(2):425–442, 2001.

[WDRP02] N. Wicker, D. Dembele, W. Raffelsberger, and O. Poch. Density of points clustering, application to transcriptomic data analysis. *Nucleic Acids Research*, 30(18):3992–4000, 2002.

[WL93] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993.

[WLLH05] B. Wang, X.-F. Li, F. Liu, and F.-Q. Hu. Color text image binarization based on binary texture analysis. *Pattern Recognition Letters*, 26(11):1650–1657, 2005.

[WM97] D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[XOX01] Y. Xu, V. Olman, and D. Xu. Minimum spanning trees for gene expression data clustering. *Genome Informatics*, 12:24–33, 2001.

[XOX02] Y. Xu, V. Olman, and D. Xu. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545, 2002.

[Xu96] L. Xu. How many clusters?: A Ying-Yang machine based theory for a classical open problem in pattern recognition. In *Proceedings of the 1996 IEEE International Conference on Neural Networks,* 3-6 June, Washington, DC, USA, volume 3, pages 1546–1551. IEEE, 1996.

[Xu97] L. Xu. Bayesian Ying-Yang machine, clustering and number of clusters. *Pattern Recognition Letters*, 18(11–13):1167–1178, 1997.

[XWI05] R. Xu and D. Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[YCC00] Y. Yao, L. Chen, and Y. Q. Chen. Using cluster skeleton as prototype for data labeling. *IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics*, 30(6):895–904, 2000.

[Zah71] C. T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, C-20(1):68–86, 1971.

[Zar02] C. D. Zaroliagis. Implementations and experimental studies of dynamic graph algorithms. In *Experimental Algorithmics: From Algorithmic Design to Robust and Efficient Software*, volume 2547 of *Lecture Notes in Computer Science*, pages 229–278, Berlin Heidelberg, 2002. Springer Verlag.

[ZZZL07] C. Zhang, X. Zhang, M. Q. Zhang, and Y. Li. Neighbor number, valley seeking and clustering. *Pattern Recognition Letters*, 28(2):173–180, 2007.

Kuopio University Publications H. Business and Information technology

H 1. Pasanen, Mika. In Search of Factors Affecting SME Performance: The Case of Eastern Finland.
2003. 338 p. Acad. Diss.

H 2. Leinonen, Paula. Automation of document structure transformations.
2004. 68 p. Acad. Diss.

H 3. Kaikkonen, Virpi. Essays on the entrepreneurial process in rural micro firms.
2005. 130 p. Acad. Diss.

H 4. Honkanen, Risto. Towards Optical Communication in Parallel Computing.
2006. 80 p. Acad. Diss.

H 5. Laukkanen, Tommi. Consumer Value Drivers in Electronic Banking.
2006. 115 p. Acad. Diss.

H 6. Mykkänen, Juha. Specification of reusable integration solutions in health information systems.
2006. 88 p. Acad. Diss.